Computational Annotation of Transcription Factor Binding Sites in D. melanogaster Developmental Genes

Vipin Narang[1] **Wing-Kin Sung**[2]
vipinnar@comp.nus.edu.sg ksung@comp.nus.edu.sg

Ankush Mittal[3]
ankumfec@iitr.ernet.in

[1,2] Department of Computer Science, 3 Science Drive 2, National University of Singapore, Singapore-117543
[3] Department of Electronics and Computer Engineering, Indian Institute of Technology, Roorkee-247667, Uttaranchal, India

Abstract

Drosophila melanogaster is one of the most important organisms for studying the genetics of development. The precise regulation of genes during early development is enacted through the control of transcription. The control circuitry is hardwired in the genome as clusters of multiple transcription factor binding sites (TFBS) known as cis-regulatory modules (CRMs). A number of TFBS and CRMs have been experimentally annotated in the Drosophila genome. Currently about 661 CRM sequences are known, of which 155 have been annotated with 778 TFBS. This work attempts computational annotation of TFBS in the remaining 506 uncharacterized Drosophila CRMs. The difficulty of this task lies in the fact that experimental data is insufficient for constructing reliable positional weight matrices (PWM) to predict the TFBS. Thus a novel feature extraction and classification method for TFBS detection has been implemented in this work. The method achieves both high sensitivity and low false positive rate in cross-validation studies. As a result of this work, a new database has been compiled which aggregates all the CRM and TFBS annotation information for Drosophila available to date, and appends new TFBS annotations.

Keywords: cis-regulatory modules, transcription factor binding sites, positional weight matrix, feature selection

1 Introduction

Drosophila melanogaster (fruit fly) is an important model organism for studying the process of development. Development occurs in a series of stages, including embryogenesis, three larval stages, a pupal stage, and finally the adult stage. The early Drosophila embryo exists as a multinucleate cell called syncytial blastoderm, which gradually divides into individual cells forming the cellular blastoderm. During this phase, differential expression of genes across the embryo determines the major body axes and segment boundaries. Subsequently morphologically distinct segments and organs are formed.

The precise regulation of gene expression from fertilization till organ development is accomplished largely through transcriptional control. A number of genes expressed in the developmental phase encode transcription factors (TFs). The TFs operate in a hierarchical fashion so that TFs released at one stage lead to the expression of genes that release TFs for the next stage. At each stage the complexity of expression pattern increases. Initially the maternal-effect TFs expressed by the mother during oogenesis acquire an anterior-posterior (A-P) concentration gradient across the syncytial blastoderm embryo. These genes encode TFs that regulate the expression of the gap genes, which roughly subdivide the embryo into broad regions along the A-P axis. The gap genes encode TFs that regulate the expression of the pair-rule genes, which divide the embryo into pairs of segments. As the syncytial blastoderm stage ends and the embryo cellularizes, TF products of pair rule genes

express segment polarity genes which set the A-P axis of each segment. Finally, the TFs encoded by segmentation genes initiate a family of homeotic genes, which cause structures like legs, wings, and antennae to develop on the particular segments.

The TFs regulate gene expression by exerting their activating or repressing influence upon basal transcription. They bind to specific DNA sites in the regulatory region of the target genes in order to interact with the basal transcription apparatus. Binding sites for several TFs are often present in close proximity as a cis-regulatory module (CRM). The combinatorial activity of multiple TFs in a CRM helps to achieve precise control over both the expression level and the location (tissue) of transcription. Analysis of tissue specific regulatory sequences indicates that a CRM as a whole contributes in a specific way to the overall regulatory output. In fact, different parts of the overall regulatory task are carried out by multiple CRMs influencing the same basal transcription apparatus. The expression pattern generated by multiple CRMs is physically a sum of the patterns mediated by the individual CRMs. Thus CRMs are frequently found among Drosophila developmental genes that are expressed in complex spatial patterns and at different times, such as the even skipped gene.

For a number of Drosophila genes, the associated CRMs have been experimentally determined. Recently a comprehensive collection of over 600 experimentally determined CRMs in Drosophila was compiled in the REDfly database [5]. Supplementing to the experimental data, computational techniques have been found valuable in discovering novel CRMs [3,6-13]. A fair proportion of the computational CRM predictions have been experimentally validated and found to be accurate. Despite the success of the computational method, a limitation has been the breadth of coverage. To the best knowledge of the authors, all the existing computational prediction studies have concentrated upon gap and pair-rule genes in which the CRMs are composed of a handful of maternal and gap factors listed in Table 1. Although the computational prediction techniques are general in nature, i.e. not restricted to any specific group of TFs, their application has been limited. The main reason is that almost all computational techniques rely on positional weight matrices (PWMs) to detect transcription factor binding sites (TFBS) in genomic sequences, but for many of the TFs in Drosophila, experimental data is insufficient for constructing reliable PWMs.

Table 1: Transcription factors commonly referred to in computational CRM prediction studies.

TF	Makstein et al. (2002)	Berman et al. (2002)	Rajewsky et al. (2002)	Lifanov et al. (2003)	Schroeder et al. (2004)
Bicoid		✓	✓	✓	✓
Caudal		✓	✓		✓
Dorsal	✓		✓		
Giant				✓	✓
Hunchback		✓	✓	✓	✓
Knirps		✓	✓	✓	✓
Kruppel		✓	✓	✓	✓
Stat92E					✓
Tailless			✓		✓
TorRE			✓		✓

The coverage of computational CRM prediction can be extended by including information on more TFs. In another recent work, more than 1300 experimental TFBS annotations for over 80 different TFs in the Drosophila genome were compiled in the Drosophila DNase I Footprint Database (also known as FlyReg database) [2]. Researchers [16] have already prepared PWMs for several TFs using this TFBS data. In the present study, the accuracy and practicability of these PWMs in detecting TFBS was tested. Unfortunately it was found that most of them have low performance and are unusable. Therefore need was felt to find an alternative way of computationally annotating/detecting TFBS in the CRMs.

A novel computational method for detecting TFBS in Drosophila CRMs is developed in this paper. In a recent work, the use of hexamer strings as features was reported as effective in distinguishing between CRM and non-CRM sequences in Drosophila [4]. Hexamers that were overrepresented in CRM sequences as compared to non-CRM sequences were extracted as features for the classification task. In the present problem, however, overrepresented hexamer string features did not perform well for TFBS detection. However building in the same direction, a statistical method of extracting string features that distinguish TFBS from rest of the regulatory sequence regions has been derived. Using these features a classifier is built to distinguish TFBS from non-TFBS sequences. Cross-validation studies show that this approach has a reliable performance in detecting TFBS, which could not be possible using PWMs.

As a result of the present study, a new database has been compiled which includes all the CRM and TFBS

annotation information available to date, as well as the results of the present study. It is intended to provide to the research community an extensive and reliable annotation of Drosophila CRMs and their TFBS composition.

2 Data

2.1 Collection of data

The Drosophila genome has more than 165 million bases in four pairs of chromosomes, containing an estimated 14,140 genes. A total of 235 Drosophila genes were selected in the present study from three different resources: (a) Drosophila DNase I Footprint Database v2.0 (FlyReg database) [2], (b) computational cis-regulatory module predictions by Schroeder et al. (2004) [13], and (c) REDfly database [5]. The overlap among the three data sources is shown in Figure 1.

For 85 genes, experimental annotations of 1066 TFBS for 83 known transcription factors were collected from the FlyReg Database. This is a subset of the FlyReg database, leaving out entries with unknown transcription factor or gene information. For 196 genes, a total of 619 experimentally annotated CRMs were obtained from the REDfly database. The FlyReg and REDfly databases had 52 genes in common, so that both experimentally annotated TFBS and CRMs could be obtained for these genes. Interestingly, the annotated TFBS overlapped the annotated CRM regions for all genes except one. There were thus 778 known TFBS falling within 155 known CRMs across 51 genes. These genes comprised the training set in this study since extensive annotation was available for them.

For the rest 184 genes, only partial information of either TFBS or CRM annotations was available. These genes formed the subject of our attempted extension. The study of Schroeder et al. (2004) added information of 42 additional CRMs (3 experimental and 39 predicted), making the total number of available CRMs as 661. However none of these CRMs overlapped any of the known TFBS from the FlyReg database, and so they did not contribute towards the training set.

2.2 Preparation of training and test datasets

As described above, the experimentally well annotated set of 51 genes with 155 CRMs and 778 TFBS is used as the training dataset. These CRM sequences are of length about 172 kilobases. They are partitioned into three different types of sequence segments: (1) CRM-TFBS segments, (2) CRM-non-TFBS segments, and (3) non-CRM sequences. The 778 TFBS falling within the 155 CRMs form the set of CRM-TFBS segments.

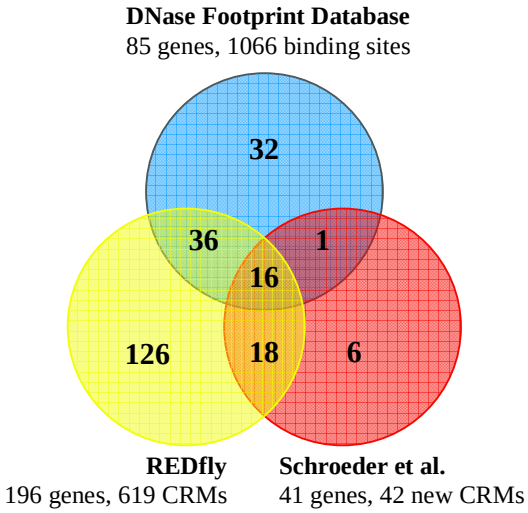

Figure 1. Data acquired from different resources.

These TFBS vary in length from 5bp to 140 bp. TFBS for even the same TF have different lengths due to the nature of the experimental annotation technique. The CRM segments other than TFBS have been considered as CRM-non-TFBS segments. It is important to note however that the TFBS annotations obtained from RedFly database may not cover all the TFBS present in the CRMs. Thus it is expected that the CRM-non-TFBS segments will still have some remaining TFBS within them.

The non-CRM sequences are any genomic region other than CRMs. A caution is required in collecting non-CRM sequences because most of the genes in Drosophila are narrowly spaced with an inter-distance of only about a few hundred bases. The regulatory region of one gene frequently overlaps the adjacent genes. Thus only well confirmed sequence regions must be selected as non-CRM sequences. In this study, 100 non-CRM segments present in the middle to two adjacent CRMs in well annotated genomic regions have been obtained. They span a total of 1,046 kilobases. It should also be noted that TFBS may be present within the non-CRM sequences since they are not restricted to lie only within CRMs.

2.3 Objective of the study

The present study seeks to computationally annotate the binding sites within 506 CRMs for which experimental annotation is not available, while using the set of 155 CRMs with 778 experimentally annotated binding sites as the training data. The data compiled from the abovementioned resources, as well as the new information contributed in this study, have been consolidated as Drosophila Cis-Regulatory Database (DCRD), which is available at the website http://www.comp.nus.edu.sg/~bioinfo/Drosophila.

3 Positional Weight Matrix for TFBS Detection

Positional weight matrix (PWM) is the most common way of representing and detecting TFBS for a particular TF [14,15]. It records the preference of nucleotides at each TFBS position in a $4 \times N$ matrix. The entries of the matrix are the frequencies, $f_{b,i}$, of the four nucleotides, $b \in \{A,C,G,T\}$, in positions, $i \in \{1,2,\ldots,N\}$, among all the TFBS obtained for a particular TF. Here N is the length of each TFBS. The PWM is used to detect TFBS in a given uncharacterized sequence as follows. At each position, p, in the uncharacterized sequence, a window of length N = number of columns in the PWM is selected. Let the currently selected sequence window be denoted by $S = S_1 S_2 \ldots S_N$, $S_i \in \{A,C,G,T\}$. The "matrix score" for this window is calculated as:

$$\text{Matrix score} = \frac{\sum_{i=1}^{N} \ln\left(f_{S_i,i}\right) - \sum_{i=1}^{N} \ln\left(f_{\min_i,i}\right)}{\sum_{i=1}^{N} \ln\left(f_{\max_i,i}\right) - \sum_{i=1}^{N} \ln\left(f_{\min_i,i}\right)}$$

where $\max_i$ and $\min_i$ represent the rows for which $f_{b,i}$ is maximum and minimum respectively in the column i. The matrix score is a real number within the range (0,1). If the matrix score for the window S exceeds a chosen threshold value, then it is marked as a potential TFBS. The TFBS detection performance of a PWM can ideally be quite high, but in practice it varies depending upon the number and quality of TFBS used to learn the PWM and the selected score threshold.

The TFBS available in the FlyReg database have been used by the research community to learn PWMs for 75 different TFs [16]. Learning the PWMs was difficult because of two reasons. Firstly, the TFBS were of varying lengths and were not aligned relative to each other. Secondly, the number of TFBS available was too little for most TFs. As shown in Figure 2(a), less than 10 TFBS were available for 44 TFs, whereas sufficient number of TFBS (30 or more) were available only for 11 TFs. The PWM based motif finding tool MEME [1] was used to align the TFBS and learn PWMs of specified lengths.

To test the performance, all 75 PWMs were used to detect TFBS in the 155 training CRM sequences. For each PWM, several different matrix score thresholds in the range (0,1) were tried. The quality of TFBS detection was measured in three aspects – sensitivity, specificity, and correlation coefficient, which are defined as follows:

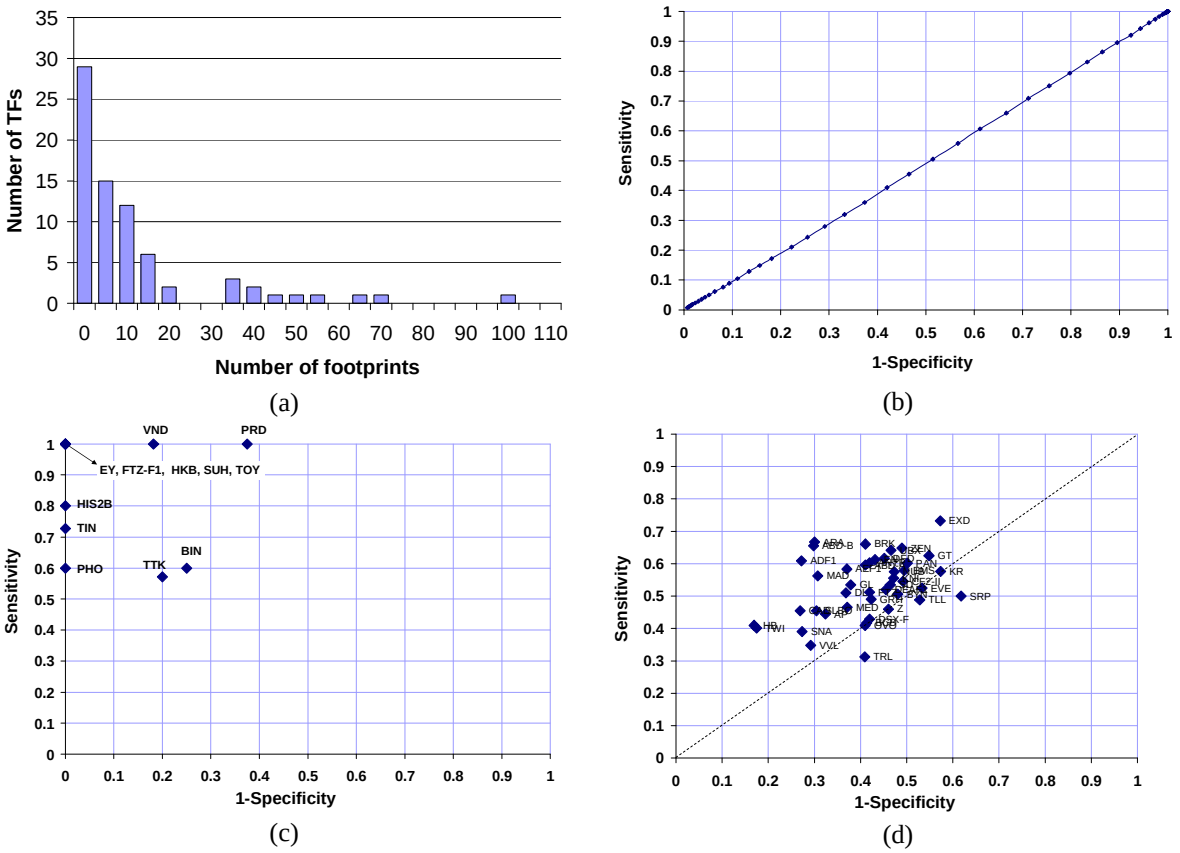

Figure 2. TFBS detection performance of PWM: (a) histogram showing the number of TFBS (footprints) available per TF in the FlyReg database, which were used to construct the matrices, (b) ROC curve for combined prediction accuracy of all matrices, (c) performance of 13 best matrices at their respective best chosen thresholds, (d) performance of the remaining matrices at their respective best chosen thresholds.

$$Sensitivity = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP}, \quad CC = \frac{TP \times TN - FP \times FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)},$$

where,

TP = predicted TFBS overlaps actual TFBS,
FP = predicted TFBS overlaps actual non-TFBS,
TN = predicted non-TFBS overlaps actual non-TFBS,
FN = predicted non-TFBS overlaps actual TFBS

In the physical sense, sensitivity refers to the percentage of actual TFBS that could be successfully predicted, specificity refers to the percentage of actual non-TFBS that could be successfully rejected, and correlation coefficient measures the difference between number of correct and incorrect predictions on a scale of –1 to 1. The overall classification performance is shown through the receiver-operator characteristics (ROC curve). The degree of accuracy of detecting TFBS and rejecting non-TFBS is seen in how much the ROC curve deviates from the diagonal.

Figure 2(b) shows the ROC of combined TFBS detection performance of the 75 PWMs over the 155 CRM sequences. The near 45 degree slope of ROC curve indicates that the overall performance in classifying TFBS vs. non-TFBS is almost nil on this dataset. However, caution is necessary in interpreting this result since each PWM finds TFBS for a specific TF and the best performance threshold will very for each PWM. Individual prediction performance of thirteen best PWMs at their respective best chosen threshold values is shown in Figure

2(c). These matrices perform well in terms of high sensitivity and low false positive rate, but the number of TFBS associated with these matrices represents a very small fraction (6%) of the total number of TFBS. For majority of the TFBS, the respective PWMs have very low performance even on an individual basis as shown in Figure 2(d). Thus these PWMs are not reliable towards the annotation and detection of TFBS in Drosophila CRMs.

4 Feature Extraction for TFBS Detection

Owing to the unsatisfactory performance of the PWMs, an alternative approach was sought to reliably annotate TFBS in the Drosophila CRMs. In this work a scheme of feature extraction, including (i) feature selection, (ii) weighting, and (iii) classification has been developed to obtain better TFBS detection performance. *The features in this study are strings of any length which are potentially useful towards discriminating TFBS and non-TFBS segments*. The feature extraction scheme is conceptually explained in Section 4.1, and then specific details of how the parameters are tuned for best performance are provided in Section 4.2.

4.1 Overview of the feature extraction scheme

The goal of feature extraction is to obtain an as small as possible set of relevant features which can be used to distinguish between TFBS and non-TFBS segments. Strings of lengths 2 to 8 were tested as possible features. The chi-square statistic was used to measure the relevance of a feature towards distinguishing TFBS and non-TFBS segments in CRMs. The chi-square statistic is computed as follows. For any length-*k* string, *f*, (where $k \in \{2,3,\ldots,8\}$) the number of its occurrences in TFBS and non-TFBS CRM sequences are counted and a 2×2 contingency table is formed as follows:

a = number of occurrences of *f* in TFBS sequences	*b* = number of occurrences of all length-*k* strings other than *f* in TFBS sequences
c = number of occurrences of *f* in non-TFBS sequences	*d* = number of occurrences of all length-*k* strings other than *f* in non-TFBS sequences

The chi-square statistic for the string, *f*, is then given by the formula

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}, \quad P-value = \int_{\chi^2}^{\infty} f(x,df)\,dx,$$

where $f(x,df)$ is the chi-square probability density function with *df* being the number of degrees of freedom, which is equal to 1 in the present case. All features with chi-square P-value lower than a fixed cutoff were selected. P-value cutoffs of 0.05 or 0.01 are recommended in standard practice.

The selected features were then weighted according to their contribution in classifying between CRM-TFBS and CRM-non-TFBS segments. Four different weighting schemes were tried as shown in Table 2. All schemes in Table 2, except for binary, give a positive value to the weight regardless of whether the feature favors the TFBS or the non-TFBS class. This is unsuited for classification; therefore the weights have been assigned a positive or negative sign based upon the class they favor. The sign is according to the sign of the correlation coefficient, ϕ, of a feature, which is given as

$$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Next the weighted features were used for the detection of TFBS in an uncharacterized sequence as follows. At any position, *p*, in the uncharacterized sequence, a short window of length *l* is selected. Then all the occurrences of selected features in this window are obtained, and their weights are summed together. If the total weight exceeds a certain threshold, then the window is classified as a TFBS.

Table 2: Typical feature weighting schemes.

Weighting scheme	Formula
Chi square (CHI)	$w = \frac{(ad-bc)^2 (a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$
Probability ratio (RAT)	$w = \frac{a(c+d)}{c(a+b)} + \frac{c(a+b)}{a(c+d)}$
Inverse frequency (INV)	$w = \ln\left(\frac{b+d}{a+c}\right)$
Binary (BIN)	$w = \begin{cases} +1 & \text{if } ad - bc \geq 0 \\ -1 & \text{if } ad - bc < 0 \end{cases}$

4.2 Tuning the parameters

Several combinations of feature length *k*, chi-square cutoff, weighting scheme and window length were tried to obtain the best performance of classification between TFBS and non-TFBS segments. A 10-fold cross validation was performed for each parameter set. In this procedure ten experiments are performed, where in each experiment 90% of the CRM-TFBS and CRM-non-TFBS segments are used to select and weight the features, while the rest 10% are used to test the performance. Thus each data is covered for both training and testing. Among the several combinations of parameters tried, only the representative results are shown for brevity. The effect of varying only one parameter at a given time while keeping the others constant is shown. The ROC curves showing the performance were obtained by varying the score threshold used to classify a window as TFBS.

The effect of varying the chi-square cutoff is shown in Figure 3(a). It is observed that the set of features selected with P-value cutoff of 0.05 gave the best performance of TFBS detection. A P-value cutoff of higher than 0.05 would be statistically unsound and was therefore not attempted.

The effect of varying the feature weighting scheme is shown in Figure 3(b). The probability ratio and inverse frequency weighting schemes performed better than the binary weighting scheme. The binary weighting scheme considers only the presence or the absence of a feature. Therefore the result implies that use of a weighting scheme is beneficial for classification as compared to merely considering the presence or absence of features. Chi-square, being a nonlinear function, performs the worst as a weighting function.

Figure 3(c) shows several combinations of feature lengths used. An important observation is that lower feature lengths introduce noise into the classification and therefore diminish performance. Longer feature length seem to improve performance, but this can be attributed to over-training, which is seen in the jaggedness of the ROC curve for $k = 8$ case. Beyond $k = 8$, a high degree of over-training is expected. Therefore a combination of feature lengths $k = 6$, 7 and 8 was found to be the most appropriate, having both high performance and generality.

Finally, the effect of varying the window length is observed in Figure 3(d). Initially performance improves for increasing window length, but above a threshold value of $l = 13$, the performance starts gradually diminishing.

Thus finally the following set of parameters is selected: (i) Feature length = 6, 7 and 8, (ii) Chi-square P-value cutoff = 0.05, (iii) Feature weighting scheme = Probability ratio, and (iv) Window length = 13. The classification performance for this parameter set on the dataset of 155 CRM sequences is shown in Figure 4. The threshold (marked by cross-wire in Figure 4) was selected as the one which yielded the highest value of correlation coefficient. This will be used for the annotation of TFBS in uncharacterized Drosophila CRMs.

5 TFBS Annotation in Uncharacterized Drosophila CRMs

In this section, the accuracy of the feature-based TFBS detection scheme is further validated. Then with confidence the satisfactory performance of the extracted features, annotation of TFBS is performed in the complete set of 661 CRMs, of which 506 are currently uncharacterized.

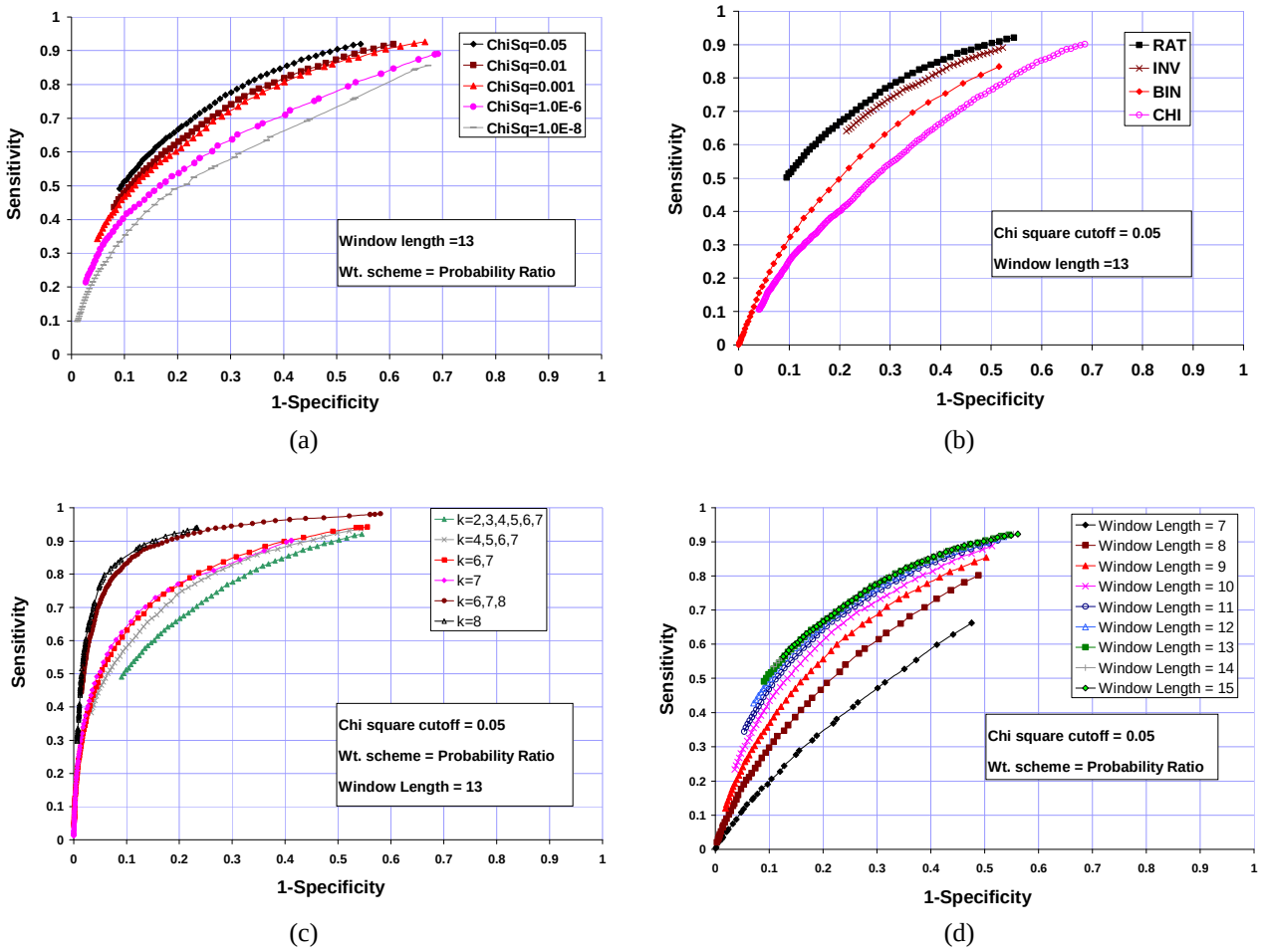

Figure 3. Effect of varying the following parameters on TFBS detection performance: (a) chi-square cutoff for feature selection, (b) feature weighting scheme, (c) feature length, (d) window length.

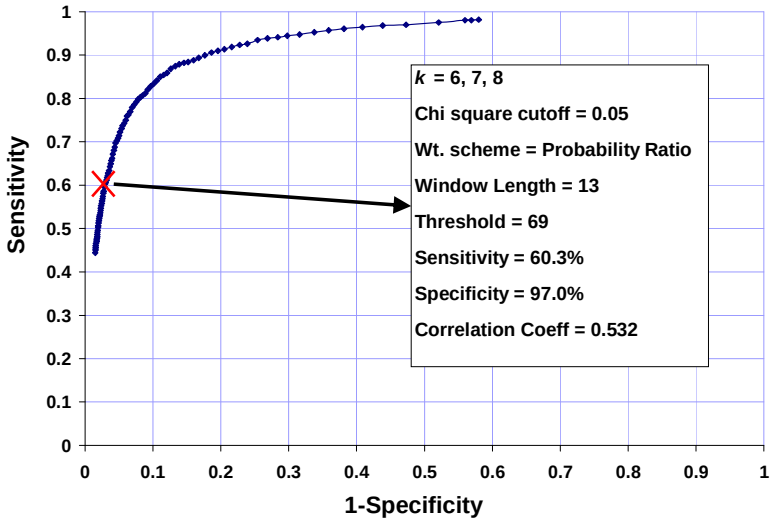

Figure 4. TFBS detection performance for the finally selected parameter set.

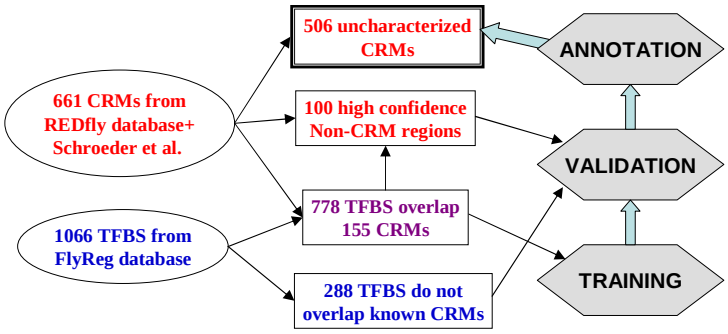

Figure 5. Extraction of datasets for training, validation and annotation.

5.1 Performance validation of feature-based TFBS detection

Figure 5 shows how the datasets for performance validation were extracted. From the available source of 661 CRMs and 1066 TFBS, an overlapping set of 155 CRMs and 778 TFBS was used for training (i.e. feature extraction and weighting) in Section 4. The rest 288 TFBS, which are not associated with any of the known CRMs, may be used for validation experiments. Also, 100 sequence regions that lie between adjacent CRMs have been extracted. Since the 155 CRM annotations used for this purpose are of high quality, the extracted sequences are non-CRMs with a good certainty. These sequences can be used as a negative dataset to test degree of false positives in TFBS detection.

In the dataset of 288 sequences containing TFBS, the feature-based scheme could detect 143 TFBS accurately (sensitivity = 49.6 %) with a high specificity of 95.4%. In the non-CRM sequences, an average of 4.9 TFBS predictions per 1000 bp of sequence was reported. This may be compared to the number of TFBS predicted in CRM regions, which are about 6.8 per 1000 bp. The high prediction accuracy of TFBS in the CRM sequences and the low false positive rate in non-CRM sequences is in support of the validity of the approach.

5.2 Annotation of uncharacterized CRMs

With sufficient confidence in the TFBS prediction accuracy of feature-based approach, the set of 506 uncharacterized CRMs was annotated. A total of 9218 predictions were made, which amounts to, on an average, 7.96 TFBS per 1000 bp of sequence. The predicted TFBS and all the supporting data have been consolidated as the Drosophila cis-regulatory database (DCRD), which is freely available at the website: http://www.comp.nus.edu.sg/~bioinfo/Drosophila.

6 Conclusions

In this study, computational annotation of TFBS was performed in 506 uncharacterized Drosophila CRMs using experimental information concerning 155 CRMs and 1066 TFBS derived from public data resources. Positional weight matrix method was found inadequate for the task as the number and quality of TFBS data was insufficient for learning reliable PWMs. A feature extraction based approach however gave good performance in both training and validation studies. The features used were strings of lengths 6, 7 and 8, and these were selected using chi-square statistics. Features were weighted using probability ratio score, and their weighted sum served as a score for classification. The study shows the effectiveness of using strings as features to model the composition of eukaryotic regulatory sequences. The results of this study may serve as a useful aid to the ongoing experimental and computational research on Drosophila developmental genetics.

References

[1] Bailey, T.L. and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings 2nd ISMB Conference, AAAI Press*, 28-36, 1994.

[2] Bergman, C.M., Carlson, J.W. and Celniker, S.E., Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster, *Bioinformatics*, 21(8): 1747-1749, 2005.

[3] Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, *Proceedings of the National Academy of Sciences of the U.S.A.*, 99(2): 757-762, 2002

[4] Chan, B.Y. and Kibler, D., Using hexamers to predict cis-regulatory motifs in Drosophila, *BMC Bioinformatics*, 6: 262, 2005

[5] Gallo, S.M., Li, L., Hu, Z. and Halfon, M.S., REDfly: a regulatory element database for Drosophila, *Bioinformatics*, 22(3): 381-383, 2006.

[6] Halfon, M.S., Grad, Y., Church, G.M. and Michelson, A.M., Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model, *Genome Research*, 12(7): 1019-1028, 2002.

[7] Lifanov, A.P., Makeev, V.J., Nazina, A.G. and Papatsenko, D.A., Homotypic regulatory clusters in Drosophila, *Genome Research*, 13(4): 579-588, 2003.

[8] Makeev, V.J., Lifanov, A.P., Nazina, A.G. and Papatsenko, D.A., Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information, *Nucleic Acids Research*, 31(20): 6016-6026, 2003.

[9] Markstein, M., Markstein, P., Markstein, V. and Levine, M.S., Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo, *Proceeding of the National Academy of Sciences of the U.S.A.*, 99(2): 763-768, 2002.

[10] Nazina, A.G. and Papatsenko, D.A., Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency, *BMC Bioinformatics*, 4:65, 2003.

[11] Papatsenko, D.A., Makeev, V.J., Lifanov, A.P., Regnier, M., Nazina, A.G. and Desplan, C., Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers, *Genome Research*, 12(3): 470-481, 2002.

[12] Rajewsky, N., Vergassola1, M., Gaul, U. and Siggia, E.D., Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo, *BMC Bioinformatics*, 3:30, 2002.

[13] Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D. and Gaul, U., Transcriptional control in the segmentation gene network of Drosophila, *PLoS Biology*, 2(9): E271, 2004.

[14] Stormo, G.D., Schneider, T.D. and Gold, L., Characterization of translational initiation sites in E. coli, *Nucleic Acids Research*, 10(9): 2971-2996, 1982

[15] Stormo, G.D., DNA binding sites: representation and discovery, *Bioinformatics*, 16(1): 16-23, 2000.

[16] http://www.flyreg.org/